LAB 12

STAT 28

April 20, 2017

Welcome to the lab 12! In this lab, you will

- Do regression diagnosis on the datasets from the previous labs.
- Learn how to do logistic regression.

Regression dianosis

Red wine dataset

Let's first look at the red wine quality prediction dataset from lab 11.

Read data.

```
wine<- read.csv("winequality-red.csv", sep = ";")
wine$quality <- wine$quality + rnorm(length(wine$quality))
```

Fit the model.

```
wine.fit <- lm(quality~volatile.acidity+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+pH+sulphates+
summary(wine.fit)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0181 -0.7802  0.0202  0.7804  4.0479
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.069106   0.752463   5.408 7.36e-08 ***
## volatile.acidity     -0.808296   0.188328  -4.292 1.88e-05 ***
## chlorides            -1.470145   0.742425  -1.980  0.04785 *
## free.sulfur.dioxide   0.003786   0.003970   0.954  0.34040
## total.sulfur.dioxide -0.003312   0.001283  -2.583  0.00989 **
## pH                   -0.517249   0.219544  -2.356  0.01859 *
## sulphates             1.245578   0.205258   6.068 1.61e-09 ***
## alcohol               0.299131   0.031367   9.537  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.21 on 1591 degrees of freedom
## Multiple R-squared:  0.144,  Adjusted R-squared:  0.1402
## F-statistic: 38.23 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Exercise 1

(a) Do regression diagnostics using the `plot` function.

```
# insert your code here to do regression diagnostics.
```

(b) Answer the following TRUE/FALSE questions based on the diagnostics plot. Uncomment your answer.

```
### I. The plot indicates heteroscedasticity.
# TRUE
# FALSE
### II. There are non-linearity between the explantory variable and response variable.
# TRUE
# FALSE
### III. The normal assumtion holds for this model.
# TRUE
# FALSE
```

(c) Identify at least two outliers from the data. (You do not need to write the code.)

```
I think the sample ??? and ??? are outliers.
```

Diamond dataset

Now let us look at the diamond dataset from lab 10.

Read the data.

```
diamonds <- read.csv("diamonds.csv")
diamonds <- diamonds[sample(1:nrow(diamonds), 1000), ]
head(diamonds)
```

```
##       carat       cut color clarity depth table price length.in.mm
## 5856   0.70 Very Good     D    VVS1  61.5    63  3920         5.78
## 47262  0.71   Premium     J    VVS2  60.3    59  1843         5.79
## 12413  1.14 Very Good     E     SI2  61.5    58  5236         6.66
## 15633  1.01 Very Good     F     VS2  63.1    60  6271         6.39
## 36797  0.34   Premium     E     VS2  62.6    58   956         4.45
## 51593  0.70 Very Good     H     VS1  61.5    58  2394         5.66
##       width.of.mm depth.in.mm
## 5856         5.64        3.51
## 47262        5.82        3.50
## 12413        6.73        4.12
## 15633        6.29        4.00
## 36797        4.43        2.78
## 51593        5.72        3.50
```

Fit a linear regression.

```
diamond.fit <- lm(price ~ carat + cut + color + clarity + depth + table, data = diamonds)
summary(diamond.fit)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     table, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9334.5  -694.9  -131.2   460.1  6264.6
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5406.16    2842.97  -1.902 0.057519 .
## carat         8738.38      92.34  94.632  < 2e-16 ***
## cutGood        796.07     264.33   3.012 0.002666 **
## cutIdeal       886.08     261.51   3.388 0.000731 ***
## cutPremium     787.22     250.57   3.142 0.001730 **
## cutVery Good   756.58     250.81   3.017 0.002623 **
## colorE        -106.70     140.13  -0.761 0.446598
## colorF         -90.10     144.45  -0.624 0.532927
## colorG        -586.64     139.80  -4.196 2.96e-05 ***
## colorH        -937.08     148.82  -6.297 4.58e-10 ***
## colorI       -1295.67     163.04  -7.947 5.23e-15 ***
## colorJ       -2176.67     220.83  -9.857  < 2e-16 ***
## clarityIF     6870.17     407.80  16.847  < 2e-16 ***
## claritySI1    4960.20     318.51  15.573  < 2e-16 ***
## claritySI2    3950.40     319.96  12.346  < 2e-16 ***
## clarityVS1    6011.68     328.77  18.286  < 2e-16 ***
## clarityVS2    5719.03     321.03  17.814  < 2e-16 ***
## clarityVVS1   6226.55     355.06  17.537  < 2e-16 ***
## clarityVVS2   6247.94     342.17  18.260  < 2e-16 ***
## depth          -17.00      31.76  -0.535 0.592662
## table          -37.51      22.87  -1.640 0.101232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1181 on 979 degrees of freedom
## Multiple R-squared:  0.9105, Adjusted R-squared:  0.9087
## F-statistic:   498 on 20 and 979 DF,  p-value: < 2.2e-16
```

Exercise 2

(a) Do regression diagnostics using the `plot` function.

```
# insert your code here to do regression diagnostics.
```

(b) Answer the following TRUE/FALSE questions based on the diagnostics plot. Uncomment your answer.

```
### I. The plot indicates heteroscedasticity.
# TRUE
# FALSE
### II. There are non-linearity between the explantory variable and response variable.
# TRUE
# FALSE
### III. The normal assumtion holds for this model.
# TRUE
# FALSE
```

Logistic regression - Customer Retension

This is a data set from IBM Watson Analytics.

This data set provides info to help you predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs.

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. They need to understand who is leaving. Imagine that you're an analyst at this company and you have to find out who is leaving and why.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Read the data.

```
retension <- read.csv("customer_retension.csv", stringsAsFactors = FALSE)
retension$SeniorCitizen[retension$SeniorCitizen == 0] = "Yes"
retension$SeniorCitizen[retension$SeniorCitizen == 1] = "No"
retension = retension[retension$MultipleLines != "No phone service", ]
retension = retension[retension$OnlineSecurity != "No internet service", ]
retension$customerID = NULL
retension$Churn[retension$Churn == "Yes"] = 1
retension$Churn[retension$Churn == "No"] = 0
retension$Churn = as.factor(retension$Churn)
retension$PhoneService = NULL
retension$PaymentMethod = as.factor(retension$PaymentMethod)

test.set = sample(nrow(retension), 500)
retension.test = retension[test.set, ]
retension = retension[-test.set, ]
```

Exercise 3

(a) Fit a logstic regression on the dataset.

```
# insert your code here to fit a logistic regression.
```

(b) Use the backward stepwise selection method to choose a model. (You can choose the criteria yourself. It can be AIC, p-value or cross-validation score.)

```
# Insert your code here to do backward stepwise selection.
```

(c) There are four payment methods available for customers.

```
levels(retension$PaymentMethod)
```

```
## [1] "Bank transfer (automatic)" "Credit card (automatic)"
## [3] "Electronic check"          "Mailed check"
```

While holding other predictors in the model constant, which of the category have the largest retention probability? (Uncomment your answer below).

```
# Bank transfer (automatic)
# Credit card (automatic)
# Electronic check
# Mailed check"
```

Which of the category have the smallest retention probability? (Uncomment your answer below).

```
# Bank transfer (automatic)
# Credit card (automatic)
```

```
# Electronic check
# Mailed check"
```

What is the probability difference between the group with largest and the smallest retention probability?

```
A. 0.3487375
B. -0.3487375
C. 0.0187330
D. -0.0187330
E. Not enough information for calculating the difference.
```

(d) Using your fitted model, make the prediction on the test set retension.test. What is your prediction accuracy? (i.e. the proportion you got right.)

```
# Insert your code here
# Hint: use `predict` function, it is just the same with doing lienar regression
# Hint: use the argument `type="response"` to get the predicted probability
# Hint: When the probality is larger than 1 we predict it as category 1, otherwise category 0
```